Statistics and Application 统计学与应用, 2025, 14(11), 121-131
Published Online November 2025 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2025.1411316

Hans汉斯

基于CNN-Transformer-BPNN组合模型的糖尿病预测研究

马一格，刘红平*

山东建筑大学理学院，山东 济南

收稿日期：2025年10月11日；录用日期：2025年11月2日；发布日期：2025年11月12日

摘 要

本文使用SMOTE方法对Kaggle公开的糖尿病数据进行过采样，并按照8:2的比例划分训练集和测试集后进行学习、建模和预测。在Transformer模型的基础上，采用双分支并行处理结构，在两侧分支分别加入CNN模型和BPNN模型，建立CNN-Transformer-BPNN组合模型，该组合模型结合了CNN的局部特征捕获能力、Transformer的全局理解能力以及BPNN的非线性映射优势，其AUC值、F2-score、精确值分别为0.9615、0.9944、0.9614，预测效果显著。本文建立的CNN-Transformer-BPNN组合预测模型可以为糖尿病早期诊察提供可靠的临床辅助，便于对病患进行及时预警和干预，对糖尿病的诊治和我国医疗系统的发展都有积极促进作用。

关键词

卷积神经网络，反向传播神经网络，分类预测，组合模型，糖尿病

Research on Diabetes Prediction Based on CNN-Transformer-BPNN Combination Model

Yige Ma, Hongping Liu*

School of Science, Shandong Jianzhu University, Jinan Shandong

Received: October 11, 2025; accepted: November 2, 2025; published: November 12, 2025

Abstract

In this paper, the SMOTE method is used to oversample the diabetes data published by Kaggle, and the

*通讯作者。

文章引用: 马一格, 刘红平. 基于 CNN-Transformer-BPNN 组合模型的糖尿病预测研究[J]. 统计学与应用, 2025, 14(11): 121-131. DOI: 10.12677/sa.2025.1411316

training set and test set are divided according to the ratio of 8:2 for learning, modeling and prediction. On the basis of the Transformer model, a CNN-Transformer-BPNN combined model is established by using a two-branch parallel processing structure and adding CNN model and BPNN model to the two branches on both sides. The combined model combines the local feature capture ability of CNN, the global understanding ability of Transformer and the nonlinear mapping advantage of BPNN. The AUC value, F2-score and accurate value are 0.9615, 0.9944 and 0.9614, respectively, and the prediction effect is remarkable. The CNN-Transformer-BPNN combined prediction model established in this paper can provide reliable clinical assistance for early diagnosis of diabetes, facilitate timely early warning and intervention for patients, and have a positive effect on the diagnosis and treatment of diabetes and the development of China's medical system.

Keywords

Convolutional Neural Network, Back Propagation Neural Network, Classification Prediction, Combination Model, Diabetes

Copyright © 2025 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

随着数字化时代的到来，机器学习等计算机技术逐渐应用到日常生活的各个方面，其中在疾病预测领域，机器学习因其对大规模数据处理的高效性愈发显得优越[1]-[11]。Bassam [12]曾利用 Logistic 回归、K 近邻算法和支持向量机三种机器学习模型，对阿拉伯地区 2 型糖尿病高风险患者在 BMI 测量基线期后 3 年、5 年和 7 年发病风险进行了预测，其中支持向量机预测效果最好。同年，Gill [13]使用了朴素贝叶斯、Logistic 回归、随机森林等 7 种模型对糖尿病患病风险进行预测，并在机器学习模型的基础上加入方差分析、遗传算法等特征选择技术进行优化，优化后的模型极大地提高了预测精度。随着集成学习技术的发展，疾病预测方法逐渐由单一的机器学习模型向随机森林、XGBoost 等集成机器学习算法过渡。杨紫森，苏津[14]等使用 XGBoost、Logistc 回归、LightGBM 等 7 种机器学习方法构建糖尿病前期风险预测模型，结果显示 XGBoost 集成机器学习模型效果最好，AUC 值为 0.662。王琦琪，戴家佳等[15]基于随机森林、GBDT 和 XGBoost 三种集成学习模型分别建立糖尿病预测模型，并与支持向量机和 BP 神经网络进行性能比较，结果显示 XGBoost 模型的准确率最高，为 97.44%，与支持向量机、BP 神经网络相比提高了 3.85%、2.57%。通过梳理现有文献发现，目前的糖尿病预测模型基本都是单一的[16]，而每种模型都不可避免或多或少地存在其自身的不足。近年来，为了弥补单一模型的遗憾，许多学者开始尝试用组合模型来建立预测模型。2023 年，祝思婷[17]对随机森林、XGBoost、LightGBM 分别进行 Voting 和 Stacking 融合，随之提出 Stacking-Voting 组合模型，但该模型的组合方式较为简单，不涉及模型内部架构的调整，且存在预测精度降低的风险。2025 年，乔松博，孙瑜，胡海[18]等在 Transformer 模型的核心层中加入 CNN 模型和 LSTM 模型，改进其内部结构，提出 CNN-Transformer-LSTM 组合模型，实现了对全国碳市场的碳排放交易价格的精准预测。

本文在乔松博，孙瑜，胡海[18]提出的 CNN-Transformer-LSTM 组合模型基础上，将 LSTM 模型替换为更适用于截面数据的 BPNN 模型，建立 CNN-Transformer-BPNN 组合模型，利用该组合模型对糖尿病患者数据进行学习、建模和预测。Transformer 模型具有良好的处理高维数据的能力，同时通过自注意力

机制有效捕捉糖尿病患者数据中各种因素之间复杂的时间和空间依赖关系，在不同时间步之间灵活地分配注意力权重，从而更好地在序列中的各个元素之间建立联系[19]；CNN 的特征提取能力优越，通过卷积层有效捕捉患者生理数据的局部模式，利用参数共享和局部感知的思想，在数据的不同区域共享权重，提高模型效率[20]；BPNN 模型相比于 LSTM 模型结构更加简单、计算效率高、过拟合风险小，更适用于不存在时间依赖关系的截面数据，同时可以处理各因素之间的复杂非线性关系和全局特征。本文建立的 CNN-Transformer-BPNN 组合预测模型可以为早期糖尿病诊察提供可靠的临床辅助，在促进医疗资源合理配置的同时，有效地降低误诊率和漏诊率，实现对高危人群的疾病预警，督促患者及时进行检查和治疗，减轻身体和经济上的负担，为糖尿病患者带来极大的便利。

2. 研究设计

2.1. 卷积神经网络

卷积神经网络(Convolutional Neural Networks, CNN)是一类包含卷积计算且具有深度结构的前馈神经网络(Feedforward Neural Networks)，是深度学习(deep learning)的代表算法之一。

卷积神经网络的结构由卷积层、池化层和全连接层组成，其对非图像数据的处理基于局部感知和权重共享机制来实现。对于一维序列数据(如时间序列数据)或多维特征向量，卷积层使用滑动窗口在数据上移动，每个卷积核专门检测特定的局部模式，比如相邻样本的特征组合，通过权重共享使同一模式检测器在整个序列上复用；池化层则对特征进行降维并增强平移不变性，保留重要特征的同时减少计算复杂度；最终全连接层整合所有抽象特征进行分类决策。这种结构能够有效捕获数据中的局部相关性，大幅减少参数数量，并保持对特征位置变化的鲁棒性，从而实现对序列数据中深层模式的层次化提取和识别。

2.2. 反向传播神经网络

反向传播神经网络(Back Propagation Neural Network, BPNN)是一种典型的多层前馈神经网络，结构包含输入层、隐藏层和输出层，各层神经元全连接。其处理数据的原理基于前向传播和误差反向传播两个阶段。第一阶段为信号的前向传播。输入数据逐层加权求和并通过激活函数进行非线性变换，最终产生输出结果。用数学公式可以表示为：

$$\begin{cases} H_j^{\mathrm{I}}\left(x_i\right)=f\left(\sum_{i=1}^{I} w_{ij}x_i+\theta_j\right), j=1,2,\cdots,J \\ H_l^{\mathrm{II}}\left(H_j^{\mathrm{I}}\right)=f\left(\sum_{j=1}^{J} w_{jl}H_j^{\mathrm{I}}+\theta_l\right), l=1,2,\cdots,L \\ O_m\left(H_l^{\mathrm{II}}\right)=\sum_{l=1}^{L} w_{lm}H_l^{\mathrm{II}}+\theta_m, m=1 \end{cases} \tag{1}$$

其中，I、J、L 分别表示输入层、隐藏层 I、隐藏层 II 的神经元数量，x_i 是输入层第 i 个神经元的数据，是隐藏层 I 第 j 个神经元的输出值，经过激活函数 f 和偏置 θ 处理后，作为隐藏层 II 的输入值。是隐藏层 II 第 l 个神经元的输出值。

第二阶段为误差的反向传播。计算输出误差并通过链式法则将误差从输出层反向传播至隐藏层，根据梯度下降算法调整各层连接权重以最小化损失函数。当输出层输出结果与实际值之间的误差较大时，要将输出层到输入层的权重 w 和偏置 θ 依次进行调整，误差小于设定值或达到最大迭代次数时，神经网络就停止运行，否则就返回前向传播阶段，数学公式表示为式(2)，

$$\begin{cases} w'_{lm} = w_{lm} - \eta \dfrac{\partial E}{\partial w_{lm}}, w'_{jl} = w_{jl} - \eta \dfrac{\partial E}{\partial w_{jl}}, w'_{ij} = w_{ij} - \eta \dfrac{\partial E}{\partial w_{ij}} \\ \theta'_m = \theta_m - \eta \dfrac{\partial E}{\partial \theta_m}, \theta'_l = \theta_l - \eta \dfrac{\partial E}{\partial \theta_l}, \theta'_j = \theta_j - \eta \dfrac{\partial E}{\partial \theta_j} \\ e_m = \dfrac{1}{2}\left(CS_{data-m} - CS_{out-m}\right)^2 \end{cases} \tag{2}$$

其中，CS_{data-m} 为输出层的结果，CS_{out-m} 为实际值，e_m 为误差。w 和 θ 分别为各层的权重和偏置。这种机制使网络能够通过多次迭代自动学习输入与输出之间的复杂映射关系，不断优化内部参数，最终实现对非线性数据的有效拟合和模式识别。

2.3. Transformer 模型

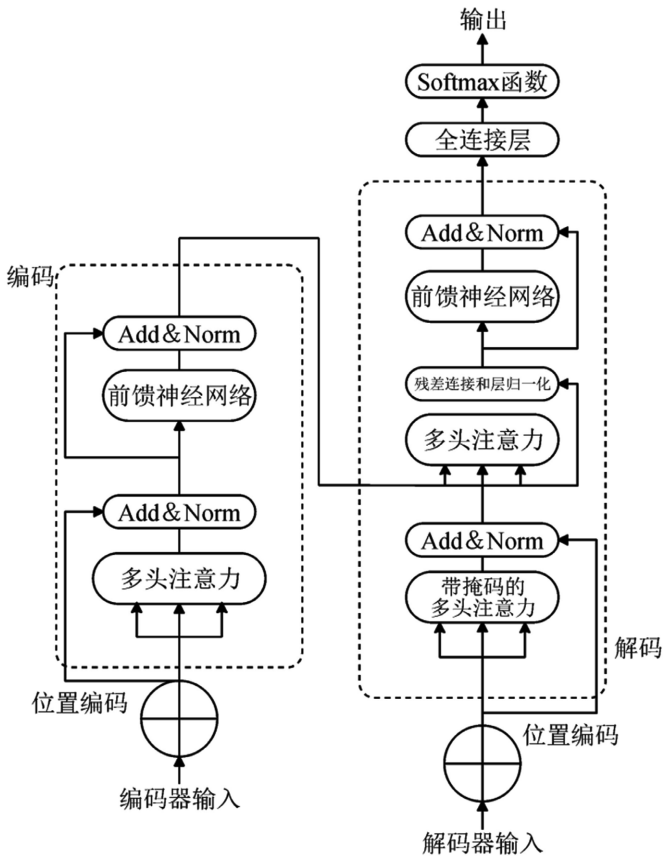

Figure 1. Transformer model structure
图 1. Transformer 模型结构

Transformer 模型一种用于自然语言处理(NLP)和其他序列到序列(sequence-to-sequence)任务的深度学习模型架构，其核心结构是基于自注意力机制(Self-Attention)和前馈神经网络组成的编码器-解码器架构，模型结构如图 1 所示。

自注意力机制可以计算输入序列中每个元素与其他所有元素的相关性权重，形成动态的特征表示，从而捕获长距离依赖关系，通过对查询(Q)、键(K)和值(V)之间的注意力分数进行计算比较和缩放处理，然后通过加权求和融合之矩阵得到最终结果，计算公式表示为：

$$\text{Attention}\left(Q,K,V\right)=\text{softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d_K}}\right)V \tag{3}$$

其中 d_K 为键向量的维度。多头注意力将模型扩展到多个表示子空间，并行学习不同类型的依赖关系，计算公式为：

$$\text{MultiHead}\left(Q,K,V\right)=\text{Concat}\left(h_{ead1},h_{ead2},\cdots,h_{eadh}\right)W_0 \tag{4}$$

式中 W_0 为权重矩阵。位置编码为序列注入顺序信息，弥补自注意力机制的位置不变性缺陷。前馈神经网络则对注意力输出进行非线性变换和特征细化。这种设计使 Transformer 能够并行处理数据，高效捕获全局上下文信息，并通过层归一化和残差连接确保训练稳定性，最终实现对数据的深度理解和表征学习。

2.4. CNN-Transformer-BPNN 组合模型

本文以 Transformer 模型为基础，采用双分支并行处理结构，两侧分支均加入 CNN 模型和 BPNN 模型，建立 CNN-Transformer-BPNN 组合模型，结构如图 2 所示。左侧分支通过 CNN 进行局部特征提取后，依次经过多头注意力机制、BPNN 模块和前馈神经网络进行深层特征学习；右侧分支同样使用 CNN 提取特征后，与左侧分支输出进行特征融合，再通过多层 Transformer 模块(包含多头注意力和前馈神经网络)进行全局依赖关系建模，最终通过全连接层输出预测结果。整个模型通过残差连接和层归一化确保训练稳定性，充分利用了 CNN 的局部特征捕获能力、Transformer 的全局上下文理解能力以及 BPNN 的非线性映射优势。

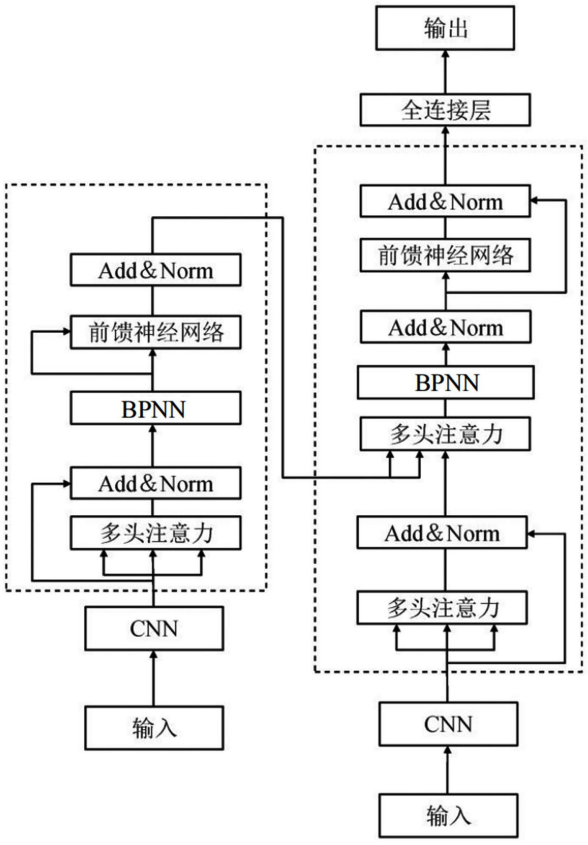

Figure 2. CNN-Transformer-BPNN combined model structure
图 2. CNN-Transformer-BPNN 组合模型结构

进行双分支设计的核心思想是功能分工与特征互补，通过构建两个结构相似但参数独立的并行网络，同时捕捉不同类型或不同来源的特征，并在不同层次进行交互或独立处理，最后融合得到更鲁棒或更全面的表示。两个分支的核心组件基本相同，但参数不共享，在训练过程中会自发地学习不同的内容。左分支通过 BPNN 学习具体的、局部化的任务特征；右分支终点为全连接层，主要学习通用的、全局化的上下文特征。双分支设计确保了模型在处理复杂任务时，既能抓住关键细节，又不失对全局的把握，从而显著提升模型的性能与泛化能力。

2.5. 模型评价指标

由于糖尿病预测需要及时捕捉患者发病情况、关注召回率，所以选择 F_2-score、准确率(accuarcy)、AUC 值作为模型的评估指标[21]。准确率和 F_β-score 的计算公式为式(5)和式(6)，计算 F_2-score 时 $\beta=2$，TP 为实际患病且预测为患病的样本数；FP 为实际未患病但预测为患病的样本数，也称误诊样本数；FN 为实际患病但预测为未患病的样本数，也叫未识别样本数；TN 为实际未患病且预测为未患病的样本数，precision 为精确率，recall 为召回率。

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{FN}+\text{TN}+\text{FP}} \tag{5}$$

$$\text{F}_\beta=\left(1+\beta^2\right)\frac{\text{precision}\cdot\text{recall}}{\left(\beta^2\cdot\text{percision}\right)+\text{recall}} \tag{6}$$

ROC 曲线横坐标为假正率(False Positive Rate, FPR)，纵坐标为查全率(True Positive Rate, TPR)，公式为式(7)、式(8)。AUC 值为 ROC 曲线下方的面积，值越大模型性能越好。

$$\text{FPR}=\frac{\text{FP}}{\text{FP}+\text{TN}} \tag{7}$$

$$\text{TPR}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{8}$$

3. CNN-Transformer-BPNN 组合模型对糖尿病患病情况进行预测

3.1. 数据选取与预处理

本文选取 kaggle 公开数据集 diabetes_prediction_dataset 作为样本进行研究，共包含 100,000 个样本，目标变量为是否患糖尿病，1 代表患糖尿病，样本量为 8500，0 代表不患糖尿病，样本量为 91,500，按照 8:2 的比例划分训练集和测试集。数据集中的“性别”、“年龄”、“是否患高血压”、“是否患心脏病”、“BMI 值”、“血压水平”、“吸烟史”、“血糖水平”8 个指标作为自变量[22]。

根据目标变量的分布可以看出患病样本仅占总体的 8.5%，存在数据不平衡的问题，会导致分类结果不准确。若对本文使用的数据集采用欠采样方法会损失大量信息，所以选择 SMOTE 过采样进行处理[23]。步骤如下：首先对于少数类中每个样本，使用欧氏距离计算其与所有其他少数类样本的距离 d，得到其 k 近邻；其次，根据样本不平衡比例设置一个采样比例以确定采样倍率 N，对每一个少数类样本，从其 k 近邻中随机选择若干个样本，假设选择的近邻为 x；最后对每个随机选出的近邻 x，分别与原样本按照式(9)构建新的样本。

$$x_{new}=x+\text{rand}(0,1)\times d \tag{9}$$

另外，原始数据集的“吸烟史”数据中存在缺失值，取值为“No Info”的样本量为 35816，缺失率较大。但是在临床上吸烟史是影响糖尿病患病的重要因素，直接删除会损失数据的有效性，所以本文选择将“No Info”视为吸烟史变量的一个取值，并使用独热编码对“吸烟史”进行处理，生成了 6 个新的二

分类变量。

预处理后的部分数据集如表 1 所示。

Table 1. Part of the data set after data preprocessing
表 1. 数据预处理之后的部分数据集

患者编号	1	2	3	4	5	6	7	8	9	10
Age (years old)	80	54	28	36	76	20	44	79	42	32
hypertension	0	0	0	0	1	0	0	0	0	0
heart_disease	1	0	0	0	1	0	0	0	0	0
BMI (kg/m^2)	25.2	27.3	27.3	23.5	20.1	27.3	19.3	23.9	33.6	27.3
HbA1c_level (%)	6.6	6.6	5.7	5	4.8	6.6	6.5	5.7	4.8	5
blood_glucose_level (mmHg)	140	80	158	155	155	85	200	85	145	100
gender_Female	1	1	0	1	0	1	1	1	0	1
gender_Male	0	0	1	0	1	0	0	0	1	0
smokinghistory_NoInfo	0	1	0	0	0	0	0	1	0	0
smokinghistory_Current	0	0	0	1	1	0	0	0	0	0
smokinghistory_Ever	0	0	0	0	0	0	0	0	0	0
smokinghistory_Former	0	0	0	0	0	0	0	0	0	0
smokinghistory_Never	1	0	1	0	0	1	1	0	1	1
smokinghistory_NotCurrent	0	0	0	0	0	0	0	0	0	0
diabetes	0	0	0	0	0	0	1	0	0	0

3.2. CNN-Transformer-BPNN 组合模型预测糖尿病

本文提出 CNN-Transformer-BPNN 组合模型对糖尿病患病风险进行预测。首先通过两个并行分支处理 15 个变量的输入数据，左侧分支使用 CNN 模块提取局部空间特征，其中包含两个卷积层，分别使用 32 个和 64 个 3 × 3 卷积核，再经过 4 头注意力机制和前馈神经网络增强特征表示；右侧分支同时使用 CNN 和 BPNN 模块分别提取特征，BPNN 中包含两个全连接层，隐藏层维度为 64。随后，模型将右侧分支的 CNN 输出、BPNN 输出以及左侧分支的最终特征进行拼接，通过一个 128 维的全连接层融合后，输入到 3 层 Transformer 编码器，每层包含 8 头注意力和 256 维前馈网络，进行全局依赖关系建模。最后，经过一个 3 层全连接输出网络得到糖尿病预测结果。整个模型共包含约 565,122 个参数，通过交叉熵损失函数和 Adam 优化器(学习率 0.001)进行端到端训练。

为了客观评估 CNN-Transformer-BPNN 组合模型的预测效果，本文另外设置了六个对照模型进行消融实验，模型 1 为 CNN-Transformer 组合模型，模型 2 为 Transformer-BPNN 组合模型，模型 3 为 CNN-BPNN 组合模型，模型 4 为单一的 CNN 模型，模型 5 为单一的 BPNN 模型，模型 6 为单一的 Transformer 模型。将本文提出的 CNN-Transformer-BPNN 组合模型与其他对照模型进行比较，各个模型的预测效果如表 2 所示。由对比结果可知，CNN-Transformer-BPNN 组合模型的准确度、F_2-score、AUC 值都是所有模型中最高的，分别为 0.9615、0.9944、0.9614，预测效果均优于对照模型，说明本文所建立的组合模型预测效果良好；另外 CNN-Transformer-BPNN 组合模型的参数量最多，训练时间最长，但单一样本预测时间较短，说明本文建立的组合模型能更加全面地捕获信息、学习数据并快速预测，在保持预测精度的

情况下又兼具高效率，避免牺牲过多性能，保持了二者的平衡。可见 CNN-Transformer-BPNN 组合模型结合了 CNN 的局部特征捕获能力、Transformer 的全局上下文理解能力以及 BPNN 的非线性映射优势，更准确地掌握了样本信息，是一种有效的预测模型。

Table 2. The experimental results of CNN-Transformer-BPNN combined model ablation
表 2. CNN-Transformer-BPNN 组合模型消融实验结果

模型	accuracy	F_2-score	AUC	模型参数量	训练时长(s)	单样本预测时间(ms)
CNN-Transformer-BPNN	0.9615	0.9944	0.9614	565122	13959.66	2.9473
CNN-Transformer	0.9579	0.9937	0.9598	237058	5759.65	2.8750
Transformer-BPNN	0.9593	0.9939	0.9593	234050	5336.34	2.5818
CNN-BPNN	0.9592	0.9938	0.9590	4454	5272.54	1.1479
CNN	0.9558	0.9934	0.9558	105794	2018.18	0.6983
BPNN	0.9521	0.9518	0.9521	9474	1521.07	0.2885
Transformer	0.9599	0.9941	0.9598	1712	5913.63	0.3707

3.3. 模型性能比较与模型解释

将 CNN-Transformer-BPNN 组合模型与现有研究所使用的集成学习模型以及融合模型进行比较。由表 3 的对比结果可知，使用 CNN-Transformer-BPNN 组合模型进行预测时模型的准确度、F_2-score、AUC 值分别为 0.9615、0.9944、0.9614，效果最好。同时与祝思婷[17]使用的 Stacking-Voting 组合模型、王琦琪等[15]使用的 BP 神经网络模型进行比较，两种模型的准确度分别为 80.20%、95.21%，F_2-score 分别为 0.7914、0.9518，AUC 值分别为 0.8020、0.9521，均低于 CNN-Transformer-BPNN 组合模型。从模型效率角度出发，CNN-Transformer-BPNN 组合模型包含 565,122 个参数，远多于基线模型，说明其能够捕捉数据中的复杂模式和关系，具有更强的泛化能力，更适用于现实场景。另外，组合模型训练时间长，但是单样本预测时间仅为 2.9473 毫秒，说明训练阶段虽然消耗了较长时间和更多算力，但模型训练后可以使用相对较少的资源进行预测，实际应用中，这种高效的资源分配在预测阶段减少了硬件需求，提升了整体系统的运行效率。

总体来看，CNN-Transformer-BPNN 组合模型在预测效果上均优于单一的集成学习模型和融合模型，说明本文所建立的组合模型精准地捕捉了患者患病的不同信息，相比于单一集成学习模型和融合模型的预测结果更准确，精度大大提高；在模型效率方面还能提供实时、低延迟的预测响应，提高了模型预测的整体效率，相比于单一集成学习模型和融合模型体现出更强的实用性和高效性。

Table 3. The results of random forest feature importance analysis
表 3. 不同糖尿病预测模型效果比较

模型	Accuracy	F2-score	AUC	模型参数量	训练时长(s)	单样本预测时间(ms)
LGBM	0.8075	0.7974	0.8074	1	0.8431	1.6902
Random Forest	0.8106	0.8011	0.8106	100	16.2009	8.3713
XGBoost	0.8851	0.8825	0.8850	1	2.6561	0.8753
BPNN	0.9521	0.9518	0.9521	9474	1521.07	0.2885
Stacking	0.8093	0.7995	0.8092	102	677.77	18.63
Stacking-Voting	0.8021	0.7914	0.8020	102	65.51	32.02
CNN-Transformer-BPNN	0.9615	0.9944	0.9614	565122	13959.66	2.9473

对患者数据进行建模之后，基于基尼指数进行随机森林特征重要性评估，定量描述特征对模型的贡献程度，分别得到不同性别群体患病的重要影响因素。基尼指数的计算公式为式(10)，

$$GI_m = \sum_{k=1}^{K} \hat{p}_{mk}\left(1-\hat{p}_{mk}\right) \tag{10}$$

K 为样本集的类别数，$\hat{p}_{mk}$ 为节点 m 样本属于第 k 类的概率估计值。变量 X_j 在节点 m 的重要性为式(11)，

$$VIM_{jm} = GI_m - GI_l - GI_r \tag{11}$$

GI_l 和 GI_r 是由节点 m 分裂的新节点的基尼指数。那么变量 X_j 在第 i 棵树上的重要性如式(12)所示，

$$VIM_{ij} = \sum_{m=1}^{M} VIM_{jm} \tag{12}$$

M 为变量 X_j 在第 i 棵树上出现的次数。对特征在每颗树上的重要性取平均值再归一化就得到了随机森林特征重要性，数值越大贡献度越高。

随机森林特征重要性分析结果如表 4 所示，从分析结果可知，影响患者患糖尿病的最重要因素为糖化血红蛋白水平、血糖水平、吸烟史、年龄，而高血压史、BMI 值、性别、心脏病史的特征重要性较低，对糖尿病患病的影响较小。现有研究中，杨海宽[24]通过主成分法对相关因素进行分析，其中性别、年龄、体重指数的特征值最高，对糖尿病的影响较为显著；张星星[25]等人分析不同肥胖指标对老年人患糖尿病风险的影响，结果显示性别、年龄、吸烟状况、高血压及 BMI 均具有统计学意义($p < 0.05$)；顾智超[26]等人使用通过 Lasso 回归筛选出年龄、性别、收缩压、舒张压、心率等因素作为预测糖尿病患病情况的关键因素；李阳[15]利用 F 检验和卡方检验得出患病重要因素前四名分别为烦渴、多尿、年龄、性别。本文选择的特征与现有研究相符，模型符合临床实际。

Table 4. The results of random forest feature importance analysis
表 4. 随机森林特征重要性分析结果

特征排序	特征名称	重要性
1	糖化血红蛋白水平	0.323807
2	血糖水平	0.200329
3	吸烟史	0.148254
4	年龄	0.113939
5	高血压史	0.068059
6	BMI 值	0.068026
7	性别	0.049271
8	心脏病史	0.028316

4. 结论

本文以 kaggle 网站的糖尿病数据集为研究对象，以 Transformer 模型为基础，采用双分支并行处理结构，两侧分支均加入 CNN 模型和 BPNN 模型，建立了 CNN-Transformer-BPNN 组合模型，结论如下。

首先，将 CNN-Transformer-BPNN 组合模型与设置的多组对照模型对比，进行消融实验，结果表明 CNN-Transformer-BPNN 组合模型的效果最好，准确度、F2-score、AUC 值分别为 0.9615、0.9944、0.9614。说明整个模型通过残差连接和层归一化确保训练稳定性，充分利用了 CNN 的局部特征捕获能力、

Transformer 的全局上下文理解能力以及 BPNN 的非线性映射优势。

其次，将 CNN-Transformer-BPNN 组合模型与集成学习模型、融合模型进行比较，结果显示相比于集成学习模型和融合模型，该组合模型大大提高了预测效果。说明本文所建立的组合模型更加精准地捕捉了患者患病的不同信息，相比于单一集成学习模型和融合模型的预测结果更准确，精度大大提高，是可靠的预测模型。另外随机森林特征重要性分析结果表明，在本文所使用的多个变量中，糖化血红蛋白水平、血糖水平、吸烟史、年龄的重要性较高，为重要因素；而高血压史、BMI 值、性别、心脏病史的特征重要性较低，对糖尿病患病的影响较小。

本文也存在局限性，虽然样本量大，但是考虑的因素较少，数据结构较单一。糖尿病的发病与遗传、生活方式、饮食方式等多方面的因素有关，特征过于单一可能会忽略各个因素之间的相互作用，从而导致模型对于某些群体的预测误差较大。另外，没有将糖尿病按照 I 型、II 型进行区分，其病因、发病机制和治疗方法存在显著差异，不对二者进行区分可能会导致模型将 I 型糖尿病患者与 II 型糖尿病患者混为一类，从而降低预测准确性，影响临床决策。后续研究可以加入免疫学标记物数据、患者的生活方式数据(如运动量、饮食偏好)、家族病史或相关基因等遗传因素进行多维度分析，同时结合糖尿病类型的细化区分来提高预测模型的准确性和适用性，从而能为临床实践提供更加精确的决策支持[27] [28]。

基金项目

国家自然科学基金青年项目(11901358)；山东省重大基础研究项目(ZR2020ZD25)。

参考文献

[1] 瞿创业, 甘立新, 乔景泉, 等. 人工智能及机器学习在骨科手术风险预测方面的作用[J]. 医学理论与实践, 2025, 38(2): 217-220.

[2] 张煊, 谢瑀, 冯亚宁, 等. 人工智能在预测肾脏疾病预后中的应用与进展[J/OL]. 中华中医药学刊: 1-12. https://link.cnki.net/urlid/21.1546.R.20250319.0954.002, 2025-04-14.

[3] 田林, 任绪泽, 涂峥程. 人工智能、机器学习和深度学习在医学诊断中的应用进展[J]. 现代医学, 2024, 52(9): 1480-1484.

[4] 王小曼, 游一鸣, 韩梦琦, 等. 基于机器学习模型对缺血性脑卒中住院期间死亡风险的预测[J]. 现代预防医学, 2024, 51(19): 3457-3462, 3482.

[5] 李雅希, 陈思平, 杨欢. 基于 Mediapipe 的脑卒中患者康复系统设计[J]. 计算机技术与发展, 2025, 35(1): 169-176.

[6] 刘忠典, 许琪, 陈伊静, 等. 心血管疾病中高风险人群颈动脉粥样硬化的识别：基于机器学习的预测模型及验证[J]. 中国全科医学, 2024, 27(30): 3763-3771.

[7] 周丽娟, 温贤秀, 吴海燕, 等. 基于机器学习算法构建慢性阻塞性肺疾病吸入剂治疗患者不良吸入风险预警模型[J]. 医药导报, 2024, 43(9): 1509-1518.

[8] 韦业, 陈广辉, 覃小伶, 等. 基于生物信息学与机器学习的坐骨神经痛与内质网应激相关生物标志物筛选及调控中药预测[J]. 中华中医药学刊, 2025, 43(7): 80-85, 287-293.

[9] 杨凯璇, 谷鸿秋. 临床预测模型常用统计模型及其 SAS 实现[J]. 中国卒中杂志, 2024, 19(5): 496-505.

[10] 李阳, 高海林, 李子杨, 等. 数据挖掘技术在糖尿病风险预测中的应用[J]. 智能计算机与应用, 2024, 14(12): 133-138.

[11] 严慧娜, 刘瑞云, 李颖, 等. 机器学习临床决策支持系统在 ICU 中应用的研究进展[J]. 护理研究, 2025, 39(7): 1199-1205.

[12] Farran, B., AlWotayan, R., Alkandari, H., Al-Abdulrazzaq, D., Channanath, A. and Thanaraj, T.A. (2019) Use of Non-Invasive Parameters and Machine-Learning Algorithms for Predicting Future Risk of Type 2 Diabetes: A Retrospective Cohort Study of Health Data from Kuwait. *Frontiers in Endocrinology*, **10**, Article 624. https://doi.org/10.3389/fendo.2019.00624

[13] Gill, S. and Pathwar, P. (2019) Prediction of Diabetes Using Various Feature Selection and Machine Learning Paradigms. In: Gunjan, V.K. and Zurada, J.M., Eds., *Modern Approaches in Machine Learning & Cognitive Science*: *A Walkthrough*,

Springer, 133-146.

[14] 杨紫森, 苏津, 唐溢乐, 等. 基于机器学习的糖尿病前期预测模型的构建及其验证[J]. 热带医学杂志, 2025, 25(5): 605-608, 620, 727.

[15] 王琦琪, 戴家佳, 崔熊卫. 基于集成学习模型的糖尿病患病风险预测研究[J]. 软件导刊, 2022, 21(4): 62-66.

[16] 叶壮. 基于机器学习方法的糖尿病预测与分析[J]. 数字技术与应用, 2024, 42(10): 33-35.

[17] 祝思婷. 基于集成学习的脑血管疾病预测研究[D]: [硕士学位论文]. 苏州: 苏州大学, 2023.

[18] 乔松博, 孙瑜, 胡海, 等. 基于 REMD-CNN-Transformer-LSTM 组合模型的碳排放交易价格预测[J]. 西安理工大学学报, 2025, 41(2): 186-196.

[19] 徐鹤, 杨丹丹, 刘思行, 等. 基于改进 Transformer 的持续血糖浓度预测模型[J]. 数据采集与处理, 2025, 40(4): 1065-1081.

[20] 吴纵凌. 糖尿病预测中不平衡数据的过采样和分类方法研究[D]: [硕士学位论文]. 西安: 西安理工大学, 2024.

[21] 陈天昕. 基于机器学习算法和深度学习算法的高炉炉温预测研究[D]: [博士学位论文]. 南昌: 江西财经大学, 2023.

[22] 张智超. 糖尿病年龄分布特点及发病年轻化相关因素研究[J]. 基层医学论坛, 2018, 22(10): 1304-1305.

[23] 刘悦. 基于机器学习的老年人抑郁症状的预测[D]: [硕士学位论文]. 济南: 山东大学, 2023.

[24] 杨海宽. 基于 GA-LightGBM 的 Stacking 模型融合的是否患有糖尿病的预测[D]: [硕士学位论文]. 武汉: 武汉轻工大学, 2023.

[25] 张星星, 张海洋, 何小菁, 等. 不同肥胖指标与老年人糖尿病患病风险的调查研究[J]. 实用老年医学, 2024, 38(9): 940-943.

[26] 顾智超, 吴昀喆, 杨帆, 等. 基于检验数据的机器学习建立 2 型糖尿病患者合并冠心病的风险预测模型[J]. 国际检验医学杂志, 2025, 46(2): 135-140.

[27] 王炳源, 高莉, 秦露伟, 等. 河南省心脑血管疾病发病预测模型的建立与评估[J]. 疾病监测, 2023, 38(10): 1239-1246.

[28] 黄丽红, 魏永越, 沈思鹏, 等. 常见新型冠状病毒肺炎疫情预测方法及其评价[J]. 中国卫生统计, 2020, 37(3): 322-326.